Lecture 14: Association and Regression

Carolyn Anderson
Department of Educational Psychology
University of Illinois at Urbana-Champaign
Topic 1: How Can We Explore the Association between Two Quantitative Variables?
Key Points

1. Identify variable type: Response or Explanatory
2. Define Association
3. Constructing scatterplots
4. Interpreting a scatterplot
5. Correlation coefficient, $r$
6. Calculating correlation coefficient, $r$
Response and Explanatory Variables

- **Response variable** (Dependent Variable) is the outcome variable on which comparisons are made.
- **Explanatory variable** (Independent variable) defines the groups to be compared with respect to values on the response variable.
- Example: Response/Explanatory
  - Blood alcohol level/# of beers consumed
  - Grade on test/Amount of study time
  - Yield of corn per bushel/Amount of rainfall
In some situations variables are not specified as “Response” and “Explanatory”.

We may just want to know about the association between them, e.g.,

- Academic achievement on different subjects.
- Relationship between different personality measures.
A personal trainer decides to track the amount of time each of her clients spends exercising and their amount of weight loss in a two month period. What are the explanatory and response variables?

a) **Explanatory variable**: time spent exercising
   **Response variable**: weight loss

b) **Explanatory variable**: weight loss
   **Response variable**: time spent exercising

c) Cannot be determined or None.
Association

- The main purpose of data analysis with two variables is to investigate whether there is an association and to describe that association.
- An association exists between two variables if a particular value for one variable is more likely to occur with certain values of the other variable.
Scatterplot

- Graphical display of relationship between two quantitative variables:
  - Horizontal Axis: *Explanatory variable*, x
  - Vertical Axis: *Response variable*, y
# Internet Usage and Gross National Product (GDP) Data Set

<table>
<thead>
<tr>
<th>Country</th>
<th>Internet</th>
<th>GDP</th>
<th>Country</th>
<th>Internet</th>
<th>GDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algeria</td>
<td>0.65</td>
<td>6.09</td>
<td>Japan</td>
<td>38.42</td>
<td>25.13</td>
</tr>
<tr>
<td>Argentina</td>
<td>10.08</td>
<td>11.32</td>
<td>Malaysia</td>
<td>27.31</td>
<td>8.75</td>
</tr>
<tr>
<td>Australia</td>
<td>37.14</td>
<td>25.37</td>
<td>Mexico</td>
<td>3.62</td>
<td>8.43</td>
</tr>
<tr>
<td>Austria</td>
<td>38.7</td>
<td>26.73</td>
<td>Netherlands</td>
<td>49.05</td>
<td>27.19</td>
</tr>
<tr>
<td>Belgium</td>
<td>31.04</td>
<td>25.52</td>
<td>New Zealand</td>
<td>46.12</td>
<td>19.16</td>
</tr>
<tr>
<td>Brazil</td>
<td>4.66</td>
<td>7.36</td>
<td>Nigeria</td>
<td>0.1</td>
<td>0.85</td>
</tr>
<tr>
<td>Canada</td>
<td>46.66</td>
<td>27.13</td>
<td>Norway</td>
<td>46.38</td>
<td>29.62</td>
</tr>
<tr>
<td>Chile</td>
<td>20.14</td>
<td>9.19</td>
<td>Pakistan</td>
<td>0.34</td>
<td>1.89</td>
</tr>
<tr>
<td>China</td>
<td>2.57</td>
<td>4.02</td>
<td>Philippines</td>
<td>2.56</td>
<td>3.84</td>
</tr>
<tr>
<td>Denmark</td>
<td>42.95</td>
<td>29</td>
<td>Russia</td>
<td>2.93</td>
<td>7.1</td>
</tr>
<tr>
<td>Egypt</td>
<td>0.93</td>
<td>3.52</td>
<td>Saudi Arabia</td>
<td>1.34</td>
<td>13.33</td>
</tr>
<tr>
<td>Finland</td>
<td>43.03</td>
<td>24.43</td>
<td>South Africa</td>
<td>6.49</td>
<td>11.29</td>
</tr>
<tr>
<td>France</td>
<td>26.38</td>
<td>23.99</td>
<td>Spain</td>
<td>18.27</td>
<td>20.15</td>
</tr>
<tr>
<td>Germany</td>
<td>37.36</td>
<td>25.35</td>
<td>Sweden</td>
<td>51.63</td>
<td>24.18</td>
</tr>
<tr>
<td>Greece</td>
<td>13.21</td>
<td>17.44</td>
<td>Switzerland</td>
<td>30.7</td>
<td>28.1</td>
</tr>
<tr>
<td>India</td>
<td>0.68</td>
<td>2.84</td>
<td>Turkey</td>
<td>6.04</td>
<td>5.89</td>
</tr>
<tr>
<td>Iran</td>
<td>1.56</td>
<td>6</td>
<td>United Kingdom</td>
<td>32.96</td>
<td>24.16</td>
</tr>
<tr>
<td>Ireland</td>
<td>23.31</td>
<td>32.41</td>
<td>United States</td>
<td>50.15</td>
<td>34.32</td>
</tr>
<tr>
<td>Israel</td>
<td>27.66</td>
<td>19.79</td>
<td>Vietnam</td>
<td>1.24</td>
<td>2.07</td>
</tr>
<tr>
<td>Yemen</td>
<td>0.09</td>
<td>0.79</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Interpreting Scatterplots

- You can describe the overall pattern of a scatterplot by the trend, direction, and strength of the relationship between the two variables
  - Trend: linear, curved, clusters, no pattern
  - Direction: positive, negative, no direction
  - Strength: how closely the points fit the trend
- Also look for outliers from the overall trend
Example: An outlier from the overall trend
Interpreting Scatterplots: Direction/Association

- Two quantitative variables $x$ and $y$ are
  - *Positively associated* when
    - High values of $x$ tend to occur with high values of $y$
    - Low values of $x$ tend to occur with low values of $y$
  - *Negatively associated* when high values of one variable tend to pair with low values of the other variable

**IN WORDS**

- Positive association: As $x$ goes up, $y$ tends to go up.
- Negative association: As $x$ goes up, $y$ tends to go down.
Example: 100 cars on the lot of a used-car dealership

Would you expect a positive association, a negative association or no association between the age of the car and the mileage on the odometer?

a) Positive association
b) Negative association
c) No association
Linear Correlation Coefficient, \( r \)

- Measures the **strength** and **direction** of the **linear** association between \( x \) and \( y \)
  - A positive \( r \) value indicates a positive association
  - A negative \( r \) value indicates a negative association
  - An \( r \) value close to +1 or -1 indicates a strong linear association
  - An \( r \) value close to 0 indicates a weak association

\[
 r = 1 - \frac{1}{n-1} \frac{\sum (x - \bar{x})(y - \bar{y})}{s_x s_y}
\]
Calculating Linear Correlation Coefficient

\[ r = \frac{1}{n-1} \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2} \sqrt{\sum(y - \bar{y})^2}} \]

<table>
<thead>
<tr>
<th>X: number of hours spent on self-study per week</th>
<th>X</th>
<th>x</th>
<th>( x )</th>
<th>Y</th>
<th>y</th>
<th>(x - ( \bar{x} ))(y - ( \bar{y} ))</th>
<th>(x - ( \bar{x} ))(y - ( \bar{y} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1 - 3 = -2</td>
<td>75</td>
<td>75 - 85 = -10</td>
<td>(-2) * (-10) = 20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2 - 3 = -1</td>
<td>79</td>
<td>79 - 85 = -6</td>
<td>(-1) * (-6) = 6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3 - 3 = 0</td>
<td>85</td>
<td>85 - 85 = 0</td>
<td>0 * 0 = 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3 - 3 = 0</td>
<td>80</td>
<td>80 - 85 = -5</td>
<td>0 * (-5) = 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3 - 3 = 0</td>
<td>89</td>
<td>89 - 85 = 4</td>
<td>0 * 4 = 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4 - 3 = 1</td>
<td>97</td>
<td>97 - 85 = 12</td>
<td>1 * 12 = 12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5 - 3 = 2</td>
<td>90</td>
<td>90 - 85 = 5</td>
<td>2 * 5 = 10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \bar{x} = 3 )</td>
<td></td>
<td></td>
<td>( \bar{y} = 85 )</td>
<td></td>
<td></td>
<td>Sum = 48</td>
<td>Sum = 48</td>
</tr>
</tbody>
</table>

\( (x - \bar{x})(y - \bar{y}) = 48 \)
\( s_x = 1.2910 \)
\( s_y = 7.5939 \)
\( n = 7 \)

\[ r = \frac{1}{7} \frac{48}{1.2910 * 7.5939} = 0.82 \]
Calculating the Standard Deviation

\[ s = \sqrt{\frac{1}{n \left( \frac{1}{n} \right)} \sum_{i=1}^{n} (x_i - \bar{x})^2} \]

<table>
<thead>
<tr>
<th>Observations</th>
<th>Deviations</th>
<th>Squared deviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_i )</td>
<td>( x_i - \bar{x} )</td>
<td>( (x_i - \bar{x})^2 )</td>
</tr>
<tr>
<td>1792</td>
<td>1792 - 1600 = 192</td>
<td>(192)^2 = 36,864</td>
</tr>
<tr>
<td>1666</td>
<td>1666 - 1600 = 66</td>
<td>(66)^2 = 4,356</td>
</tr>
<tr>
<td>1362</td>
<td>1362 - 1600 = -238</td>
<td>(-238)^2 = 56,644</td>
</tr>
<tr>
<td>1614</td>
<td>1614 - 1600 = 14</td>
<td>(14)^2 = 196</td>
</tr>
<tr>
<td>1460</td>
<td>1460 - 1600 = -140</td>
<td>(-140)^2 = 19,600</td>
</tr>
<tr>
<td>1867</td>
<td>1867 - 1600 = 267</td>
<td>(267)^2 = 71,289</td>
</tr>
<tr>
<td>1439</td>
<td>1439 - 1600 = -161</td>
<td>(-161)^2 = 25,921</td>
</tr>
<tr>
<td></td>
<td>sum = 0</td>
<td>sum = 214,870</td>
</tr>
</tbody>
</table>

\[ s^2 = \frac{214,870}{7 - 1} = 35,811.67 \quad s = \sqrt{35,811.67} = 189.24 \text{ calories} \]
Correlation coefficient on scatter-plots: Strength & Direction
Properties of Correlation

- Always falls between $-1$ and $+1$
- Correlation has a unitless measure - does not depend on the variables’ units
- Two variables have the same correlation no matter which is treated as the response variable
- Correlation is not resistant to outliers
- Only valid for linear relationship
The scatterplot below shows a graph of median family incomes per state as determined by the US Census for 1969 versus 1989. In general, what can be said about the straight line association between the variables?

a) Strong, negative  
b) Strong, positive  
c) Weak, negative  
d) Weak, positive
What value of “r” below best describes the scatterplot below?

a)  0.9  
b)  –0.9  
c)  0.3  
d)  –0.3  
e)  0
What value of “r” below would you expect if you were comparing the height and weight of all students in UIUC?

a) 0.99
b) 0.70
c) 0.00
d) – 0.70
Key Points Revisited

1. Identify variable type: Response or Explanatory
2. Define Association
3. Constructing scatterplots
4. Interpreting a scatterplot
5. Correlation coefficient, $r$
6. Calculating correlation coefficient, $r$
Addendum: Correlation Matrix

- When there is no response or explanatory variable but interest is on whether variables are linearly related
- E.g., High School and Beyond data Set of 600 high school seniors:

<table>
<thead>
<tr>
<th></th>
<th>Reading</th>
<th>Writing</th>
<th>Math</th>
<th>Science</th>
<th>Civics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>1.00</td>
<td>0.63</td>
<td>0.68</td>
<td>0.69</td>
<td>0.59</td>
</tr>
<tr>
<td>Writing</td>
<td>0.63</td>
<td>1.00</td>
<td>0.63</td>
<td>0.57</td>
<td>0.59</td>
</tr>
<tr>
<td>Math</td>
<td>0.68</td>
<td>0.63</td>
<td>1.00</td>
<td>0.65</td>
<td>0.53</td>
</tr>
<tr>
<td>Science</td>
<td>0.69</td>
<td>0.57</td>
<td>0.65</td>
<td>1.00</td>
<td>0.52</td>
</tr>
<tr>
<td>Civics</td>
<td>0.59</td>
<td>0.59</td>
<td>0.53</td>
<td>0.52</td>
<td>1.00</td>
</tr>
</tbody>
</table>
A picture of a correlation matrix
More Variable Correlations
Topic 2: Bivariate Linear Regression

How Can We Predict the Outcome of a Variable?
Key Points

1. Definition of a regression line
2. Use a regression equation for prediction
3. Interpret the slope and y-intercept of a regression line
4. Residuals and the least-squares method
Regression Analysis

The first step of a regression analysis is to identify the response and explanatory variables.

- We use \( y \) to denote the response variable (dependent variable).
- We use \( x \) to denote the explanatory variable (independent variable).
Regression by Example

Preparation for Quantitative GRE

![Scatter plot showing the relationship between the number of hours of college math courses and Quantitative GRE scores.](image)
Means of GRE by College Math

Preparation for Quantitative GRE
Plotting mean Quant GRE for each Number of Hours
Fit Line to Means

Preparation for Quantitative GRE
"Best Fit Line"

quantGRE = 333.987 + 17.99767*hoursmath
Overlay Regression in Scatter

QuantGRE vs HoursMath

quantGRE = 333.99 + 17.998 hoursmath

- N: 91
- Rsq: 0.6503
- AdjRsq: 0.6464
- RMSE: 69.641
Regression Line

- A regression line is a straight line that describes how the expected value of the response variable (y) changes as the explanatory variable (x) changes.

- A regression equation predicts the value of the response variable (y) for a given level of the explanatory variable (x)

\[ \hat{y} = a + bx \]
Example: How Can You Predict GRE based on Hours of College Math Taken?

- **Regression Equation:**
  \[ \hat{y} = 333.99 + 17.998 \times x \]

- \( \hat{y} \) is the expected (“predicted”) GRE and \( x \) is the hours of college math courses taken.

- Use the regression equation to predict the GRE of a person who took 8 hours of college math:

  \[ \hat{y} = 333.99 + 17.998 \times (8) = 477.974 \]
Interpreting the y-Intercept \((a)\)

- Interpretation of the intercept: the predicted value for \(y\) when \(x = 0\) is \(a\).

\[
\hat{y} = a + bx
\]

- Example: the predicted GRE of student who took 8 hours of college math. \(\hat{y} = 333.99 + 17.998x\)

- Note: the intercept may not have any interpretative value if no observations had \(x\) values near 0 (in our example, must take college math to graduate).
Interpreting the Slope ($b$)

- Interpretation of the slope: the predicted value of $y$ increases/decreases by $b$ units for a 1 unit increase in $x$.

\[ \hat{y} = a + bx \]

- Example: the predicted value for GRE increases by 17.998 points for 1 more hour of college math.

\[ \hat{y} = 333.99 + 17.998x \]
Slope Values: Positive, Negative, Equal to 0
In northern cities roads are salted to keep ice from freezing on the roadways between 0 and -9.5 °C. Suppose that a small city was trying to determine what was the average amount of salt (in tons) needed per night at certain temperatures. They found the following LSR equation:

$\hat{y} = 20,000 - 2,500x$

Interpret the slope. $\hat{y} = 20,000 - 2,500x$

a) 2,500 tons is the average decrease in the amount of salt needed for a 1 degree increase in temperature.

b) 2,500 tons is the average increase in the amount of salt needed for a 1 degree increase in temperature.

c) 20,000 is the average increase in the amount of salt needed for a 1 degree increase in temperature.
Regression Line

- At a given value of $x$, the equation:
  \[ \hat{y} = a + bx \]
  - Predicts a single value of the response variable
  - But... we should not expect all subjects at that value of $x$ to have the same value of $y$
  - Variability occurs in the $y$ values!
\[ \text{quantGRE} = 333.99 + 17.998 \text{ hoursmath} \]
Residuals

- Measures the size of the prediction errors, the vertical distance between the point and the regression line
- Each observation has a residual
- Calculation for each residual: \( y - \hat{y} \)
- A large residual indicates an unusual observation
QuantGRE vs HoursMath

quantGRE = 333.99 + 17.998 hoursmath
“Least Squares Method” Yields the Regression Line

- Residual sum of squares:

\[(residuals)^2 = (y - \hat{y})^2\]

- The least squares regression line is the line that minimizes the vertical distance between the points and their predictions, i.e., it minimizes the residual sum of squares.

- Note: the sum of the residuals about the regression line will always be zero.
Regression Formulas for $y$-Intercept and Slope

- **Slope:**
  \[ b = r \left( \frac{S_y}{S_x} \right) \]

- **Y-Intercept:**
  \[ a = \bar{y} \quad b(\bar{x}) \]

- **Regression line always passes through:**
  \[ (\bar{x}, \bar{y}) \]
Calculating the slope and y intercept for the regression line

\[ b = r \left( \frac{s_y}{s_x} \right) = 0.8064 \left( \frac{117.1159}{5.2477} \right) = 17.998 \]

\( \bar{x} = 8.5769 \)

\( \bar{y} = 488.3517 \)

\( s_x = 5.2477 \)

\( s_y = 117.1159 \)

\( r = 0.8064 \)

\[ a = \bar{y} - b(\bar{x}) = 488.3517 - 17.998(8.5769) = 333.99 \]

\( y \) intercept = 333.99
The Slope and the Correlation Coefficient

**Correlation:**
- Describes the strength of the linear association between 2 variables
- Does not change when the units of measurement change
- Does not depend upon which variable is the response and which is the explanatory
The Slope and the Correlation Coefficient

**Slope:**
- Numerical value depends on the units used to measure the variables
- Does not tell us whether the association is strong or weak
- The two variables must be identified as response and explanatory variables
- The regression equation can be used to predict values of the response variable for given values of the explanatory variable
Key Points Revisited

1. Definition of a regression line
2. Use a regression equation for prediction
3. Interpret the slope and y-intercept of a regression line
4. Residuals and the least-squares method
Topic 3: What Are Some Cautions in Analyzing Association?
Key Points

1. Non-linear relationship
2. Extrapolation
3. Outliers and Influential Observations
4. Correlations does not imply causation
5. Lurking variables and confounding
Non-Linear Relationship

- If the true relationship is far from a straight line
  - The correlation coefficient \((r)\) and the regression model may not describe the story precisely.
- Always plot the data!
Extrapolation

- **Extrapolation**: Using a regression line to predict y-values for x-values outside the observed range of the data
- Riskier the farther we move from the range of the given x-values
- There is no guarantee that the relationship given by the regression equation holds outside the range of sampled x-values
Outliers and Influential Points

- A *regression outlier* is an observation that lies far away from the trend that the rest of the data follows.

- An observation is influential if:
  - Its $x$ value is relatively low or high compared to the remainder of the data.
  - The observation is a regression outlier.

- Influential observations tend to pull the regression line toward that data point and away from the rest of the data.
Outliers and Influential Points

- Impact of removing an Influential data point
Correlation does not Imply Causation

- A strong correlation between x and y means that there is a strong linear association that exists between the two variables.

- A strong correlation between x and y, does not mean that x causes y.
Data are available for all fires in Chicago last year on
\[ x = \text{number of firefighters at the fires and} \]
\[ y = \text{cost of damages due to fire}. \]

1. Would you expect the correlation to be negative, zero, or positive?

2. If the correlation is positive, does this mean that having more firefighters at a fire causes the damages to be worse? Yes or No

3. Identify a third variable that could be considered a common cause of \( x \) and \( y \):
   
   a. Distance from the fire station
   b. Intensity of the fire
Lurking Variables

- A lurking variable is a variable, usually unobserved, that influences the association between the variables of primary interest
  - Ice cream sales and drowning: lurking variable is temperature
  - Reading level and shoe size: lurking variable is age
  - Childhood obesity rate and GDP: lurking variable is time

- Lurking variables are not measured in the study but have the potential for confounding the study results.
Confounding

- When two explanatory variables are both associated with a response variable but are also associated with each other, it is difficult to determine whether either of them truly causes the response.
Confounding

- Example: An experimenter wants to determine if a particular type of diet cat food (brand A or B) tends to cause more weight loss for Abyssinian and Siamese cats. To make the experiment easier to run, he gives brand A to only Abyssinian cats and brand B to only Siamese cats. What is the true cause of the difference in weight losses?

- A table of the design:

<table>
<thead>
<tr>
<th>Cat breed</th>
<th>Brand A</th>
<th>Brand B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abyssinian</td>
<td>data</td>
<td>----</td>
</tr>
<tr>
<td>Siamese</td>
<td>----</td>
<td>data</td>
</tr>
</tbody>
</table>
Key Points Revisited

1. Non-linear relationship
2. Extrapolation
3. Outliers and Influential Observations
4. Correlations does not imply causation
5. Lurking variables
Regression in R

- math =c(8.0, 8.0, 6.5, 18.5, 6.5, ..., 17.0, 8.0)
- qgre =c(610, 500, 500, 660, ..., 710, 460 )

- data = data.frame( Length = c(3.87, 3.61, 4.33, 3.43, 3.81, 3.83, 3.46, 3.76, 3.50, 3.58, 4.19, 3.78, 3.71, 3.73, 3.78), Weight = c(4.87, 3.93, 6.46, 3.33, 4.38, 4.70, 3.50, 4.50, 3.58, 3.64, 5.90, 4.43, 4.38, 4.42, 4.25) )

- result = lm(qgre ~ math, data = data)
- summary(result)
If data are in a file (file & code on course website)

<change directory to where data are>

mgdata <- read.table("math_gre.txt", header=TRUE)
attach(mgdata)
head(mgdata)
plot(hoursmath,quantGRE)
reg.result <- lm(quantGRE ~ hoursmath) # Fit regression
summary(reg.result) # Display results
plot(hoursmath,quantGRE) # Plot data again
abline(reg.result) # Overlay reg line
Practice Questions: Simple Regression

A service firm has experienced rapid growth. Because of this growth, some of the employees who handle customer calls have had to work additional hours (overtime). The firm is concerned that over-worked employees are less productive and handle fewer calls per hour than employees who work less demanding schedules. Most employees who work the “conventional” schedule put in 30-40 hours a week, depending upon demand. The firm constructed the regression model shown next relating the number of hours worked (X) to the number of calls serviced per hour (Y) for 60 employees.
<table>
<thead>
<tr>
<th>Term</th>
<th>Parameter Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>Estimate 19.8, Std Error 1.5, t Ratio 13.16, Prob&gt;</td>
</tr>
<tr>
<td>HoursWorked</td>
<td>Estimate -0.14, Std Error 0.04, t Ratio -3.74, Prob&gt;</td>
</tr>
</tbody>
</table>

(1) Does the number of hours worked impact the number of calls handled per hour, or can we explain the declining pattern seen in the plot as a simply a random coincidence?

(2) From the fitted model how many calls on average does an employee who works a 30 hour week process?

(3) For each additional 10 hours of work, how many fewer calls are processed per hour, on average? Might the drop be as large as 2 calls per hour?

(4) How would you interpret the intercept in this fitted model?
State Expenditures and SAT Scores

- Get R code in file “Lecture_14.txt”
- Get data file “state_per_student_data.txt”
- Talk about results